CHARITY DONARS PREDICTION

RAMGOPALTUMMALA (180030385)

V.BHAVANI, Assistant professor.
Department of Computer Science

Introduction

Objective of your work

There is one charity called charity which feeds homeless people, for donations this charity sends postal mails to residents of Delhi requesting to donate for a cause, from the historical data it is evident that residents who earn more than 50 thousand dollars per annum are more likely to donate. But the charity cannot be able to figure out how to send postal mails to those who most likely to donate to charity and avoid sending postal mails to those who are most likely are not going to donate to charity so that charity can save much money

Origin of your proposal

Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret, Supervised learning is an approach to creating artificial intelligence (AI), where a computer algorithm is trained on input data that has been labeled for a particular output.

Methods

Methods and Materials

Initially process is to extract data from database. to make sure we should keep overload on source database for that we used Hadoop and Sqoop to transfer data and storing. Next step, the solution for the above problem is predicting whether the resident of Delhi is earning more or less

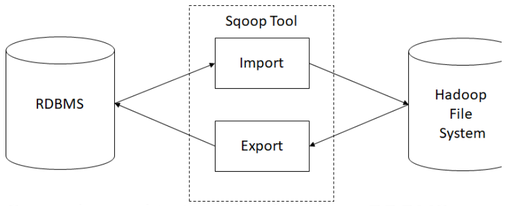
than 50 thousand dollars per annum. So, I will use a few supervised models for prediction compare

them in terms of metrics, and I will come up with the best model that suits for this problem

DecisionTreeClassifier ,KNeighborsClassifier, Naive Bayes, Logistic regression

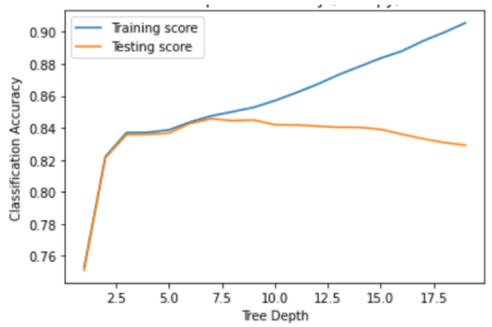
Result Analysis

Data extraction Using Sqoop



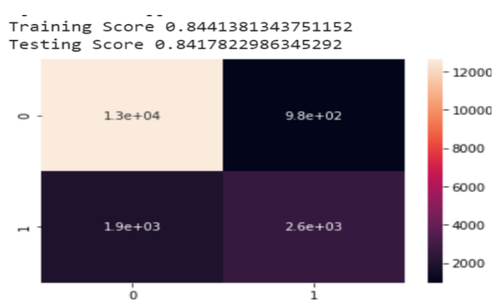
Sqoop is used to transfer data from RDBMS (relational database management system) like MySQL and Oracle to HDFS (Hadoop Distributed File System). Big Data Sqoop can also be used to transform data in Hadoop MapReduce and then export it into RDBMS

Evaluation



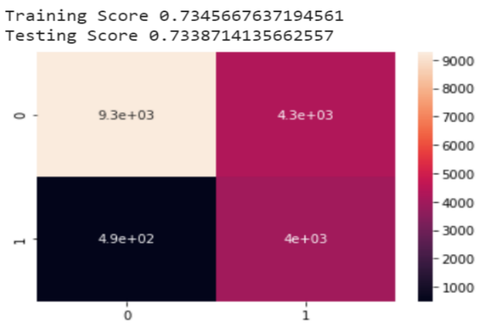
Evaluation Performance of decision tree with different parameters as follows

Logistic regression model



Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist

Naive Bayes model



Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

Results

MODEL	TRANING ACCURACY(%)	TESTING ACCURACY(%)
DECISION TREE	84.7	84.5
KNN	85.8	82.8
NAVIE BAYES	73.4	73.3
LOGISTIC REGRESSION	84.4	84.1

The best suitable model for the prediction problem is the decision tree model with tree

depth = 7, and the impurity measure is Entropy.

Conclusion

Discussion/Conclusion

Data engineering is used to make sure we should not disturb source of data so we used Sqoop and Hadoop and The best suitable model for the prediction problem is the decision tree model with tree depth = 7, and the impurity measure is Entropy.

Limitations

This Analytics can be done with less time with the help of spark which will process the data in memory. Machine learning can change any thing as days go on new algorithm can bring more accuracy.

Future Direction

Spark is used to make sure, we are making Analytics with in time so that business decisions can been taken. And also we can use kafka for making this entire data pipeline automated and scheduled.

References and Affiliations

References

Big Data Analytics A Hands-On Approach by Arshdeep Bahga, Vijay Madisetti
Big data analytics: Nada Elgendy
Logistic Regression and Reporting : Joanne Peng

Acknowledgements

Our greatest appreciation to our guide V.Bhavani, Department of Computer Science and Engineering which cannot be expressed in words for his tremendous support, encouragement and guidance for this project.